「정보시스템연구」 제28권 제4호
한국정보시스템학회
2019년 12월, pp. 271～291

http://dx.doi.org/10.5859/KAIS.2019.28.4.271

감정 딥러닝 필터를 활용한 토픽 모델링 방법론

최 병 설* · 김 남 규**

〈목 차〉

Ⅰ. 서 론

텍스트 마이닝(Text Mining)은 방대한 양의 비정형 텍스트를 문장 또는 키워드로 요약하거나, 다양한 분석에 활용하기 위한 전 단계로 정형 데이터로 구조화하기 위한 방법론이다. 텍스트 마이닝의 개념 및 기술은 4차 산업혁명으로 대변되는 빅데이터 시대에 기하급수적으로 증가하는 문서를 자동으로 요약하고 분석하기 위한 핵심적인 방법론으로 주목받고 있다. 다양한 텍스트 마이닝 응용 중 학계와 업계를 망라하여 가장 주목받는 대표적인 응용으로 토픽 모델링(Topic Modeling)을 들 수 있다.

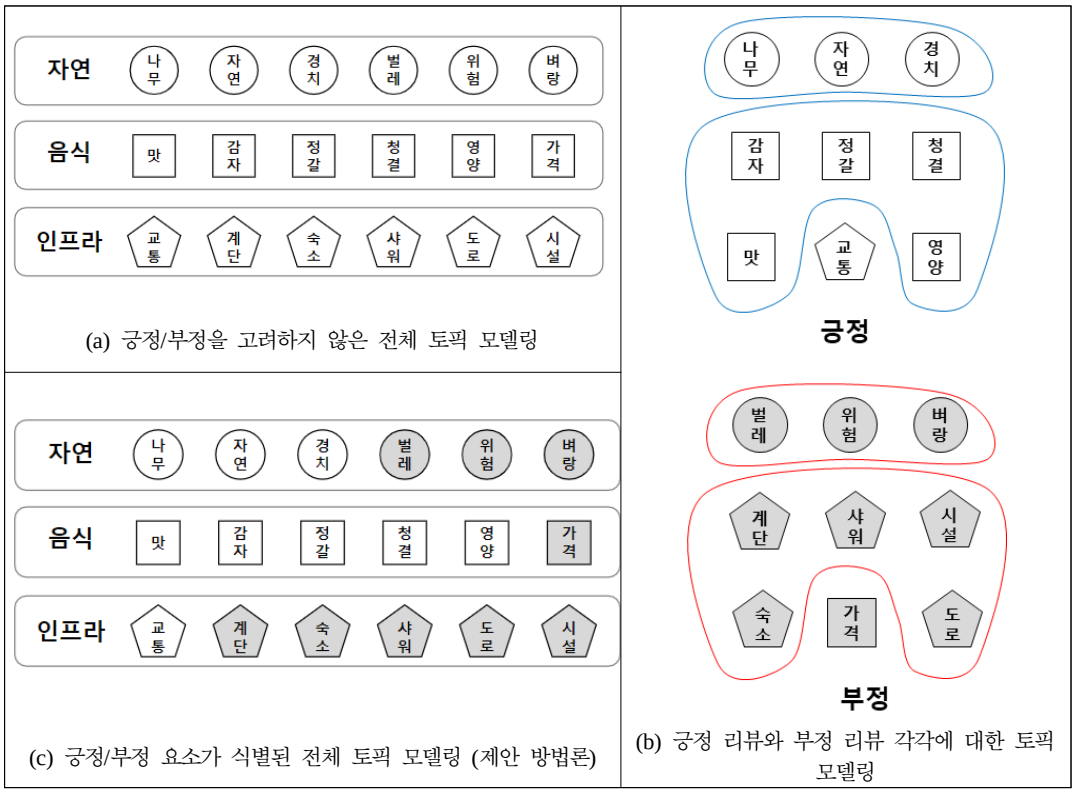
토픽 모델링은 방대한 양의 문서에 포함된 토픽을 추출하고, 각 토픽을 구성하는 키워드를 사용하여 토픽을 기술하는 일련의 과정을 의미한다. 토픽 모델링을 수행하기 위한 다양한 알고리즘이 이미 제안되었을 뿐 아니라, 이러한 최신 알고리즘을 손쉽게 적용할 수 있는 많은 상용 또는 오픈소스 소프트웨어가 폭넓게 보급되어 있다. 이로 인해 토픽 모델링은 학계뿐 아니라 다양한 산업 분야에서 활발하게 적용되고 있으며, 구체적으로 소셜 미디어에 나타난 여론 분석, 논문 및 특허 동향 분석, 뉴스 기사 분석, VOC 분석, 그리고 리뷰(Review) 분석 등에서 우수한 성과를 나타내고 있다. 특히 리뷰 분석의 경우 최근에는 여행 정보 사이트, 영화 정보 사이트, 도서 정보 사이트, 그리고 온라인 쇼핑

* 국민대학교 비즈니스 IT전문대학원, carl010@kookmin.ac.kr(주저자)
** 국민대학교 경영정보학부, ngkim@kookmin.ac.kr(교신저자)

몰의 리뷰 분석 등으로 분석 대상 컨텐츠의 주제에 맞추어 더욱 세부적으로 구체화되어 수행되는 경향을 나타낸다.

리뷰란 온라인 및 오프라인으로 제품이나 서비스를 구매한 고객들이 사용 경험이나 평가 등을 온라인상에 텍스트 형태로 기록한 데이터를 의미한다. 단순한 평점에 비해 리뷰는 훨씬 다양한 관점에서 풍부한 정보를 나타내기 때문에, 최근 고객들은 리뷰를 통해 제품이나 서비스에 대해 충분한 정보를 획득한 후 실제 구매를 실현하는 경향이 있다. 하지만 리뷰는 그 양이 방대하면서도 구조화되지 않은 형태이기 때문에, 고객이 방대한 리뷰로부터 원하는 정보를 식별해 내기 위해서는 많은 시간과 노력을 소요하게 된다. 따라서 이러한 리뷰들이 담고 있는 주요 내용을 요약하여 제시하는 과정에 토픽 모델링이 널리 활용되고 있다. 최근에는 토픽 모델링과 감정 분석을 결합하여, 리뷰가 담고 있는 감정에 따라 긍정 리뷰와 부정 리뷰를 구분하고 이를 각각 요약하여 제시하는 시도도 활발하게 이루어지고 있다.

하지만 리뷰에 대한 토픽 모델링과 감정 분석의 결합을 시도한 기존 연구들의 경우, 전술한 바와 같이 전체 리뷰에 대해 감정 분석을 우선적으로 수행한 뒤 긍정 리뷰들에 대한 토픽 모델링과 부정 리뷰들에 대한 토픽 모델링을 개별적으로 수행한다. 이러한 분석은 긍정 리뷰와 부정 리뷰 각각의 주요 토픽을 식별하기 위



<그림 1> 토픽 모델링과 감정 분석 결합 예

한 목적에는 적합하지만, 전체 리뷰를 구성하는 주요 토픽 중 긍정 요소와 부정 요소를 식별하기 위한 목적에는 부합하지 않는다. 이러한 한계는 <그림 1>의 예를 통해 설명될 수 있다.

<그림 1(a)>는 감정 분석을 고려하지 않은 토픽 모델링으로, 전체 리뷰에 대해 주요 토픽 “자연”, “음식”, “인프라”를 식별하였다. <그림 1(b)>의 경우 각 리뷰를 긍정 리뷰와 부정 리뷰로 구분한 후, 긍정 리뷰에 대한 토픽과 부정 리뷰에 대한 토픽을 각각 도출한 예이다. 토픽 모델링은 빈도에 기반을 두어 이루어지므로, 긍정 리뷰에서 낮은 빈도로 출현한 “인프라” 토픽은 긍정 리뷰를 구성하는 주요 토픽에서 누락된다. 이와 유사한 이유로 부정 리뷰에서 낮은 빈도로 출현한 “음식” 토픽은 부정 리뷰를 구성하는 주요 토픽에서 누락된다. 본 연구에서는 이러한 기존 방법론의 한계를 극복하게 위해 <그림 1(c)>와 같은 접근법, 즉 전체 리뷰에 대한 토픽 모델링을 통해 주요 토픽을 추출한 후, 각 토픽을 구성하는 긍정 요인과 부정 요인을 식별하여 기술하기 위한 방안을 제시하고자 한다.

방법론 측면에서 본 연구는 텍스트 마이닝 방법론 중 주로 활용되는 방법인 LDA나 LSA와 같이 단순히 단어의 빈도수를 기반으로 토픽 키워드를 검출하는 방법론의 단점을 보완하기 위해, 문장 내 단어와 단어의 관계를 분석하는 딥러닝(Deep Learning) 기법 중 하나인 어텐션(Attention) 메커니즘 기반의 필터를 적용하고자 한다. 특히 다양한 단어가 문장을 구성하고 이러한 문장이 모여서 긍정 또는 부정 리뷰를 형성하는 현상을 정확히 모델링하기 위해, 다양한 어텐션 메커니즘 가운데 단어 간 관계뿐 아니라 문장 간 관계까지 고려한 계층적 어텐션 (Hierarchical Attention) 메커니즘을 활용하고자 한다. 즉 토픽 모델링과 별도로 계층적 어텐션 기반의 긍정/부정 분류를 수행한 후, 토픽 모델링의 결과로 도출된 토픽 키워드에 딥러닝 기반의 긍정/부정 필터를 적용하고자 한다. 이러한 과정을 통해 전체 리뷰 관점에서의 주요 토픽을 누락 없이 표현할 수 있을 뿐 아니라, 이들 토픽을 구성하는 긍정 및 부정 요소를 명확히 식별할 수 있을 것으로 기대한다.

이후 논문의 구성은 다음과 같다. 이어지는 2장에서는 토픽 모델링과 텍스트 마이닝의 개념을 소개하기 위한 기존 연구를 살펴보며, 이와 관련하여 최근까지 수행된 연구 성과를 소개한다. 3장에서는 본 연구에서 제안하는 새로운 방법론을 개념적 설명과 예시를 통해 서술한다. 4장에서는 제안된 방법론을 검증하기 위한 실험을 수행하고, 실험에 사용된 데이터, 실험 설계 및 그 결과를 제시한다. 마지막으로 5장에서는 본 연구의 전체적인 내용과 의의를 요약한 후 본 연구의 한계를 제시하고, 이에 따른 향후 연구 방향을 제시한다.

Ⅱ. 관련 연구

본 장에서는 제안 방법론과 관련된 기존 연구 및 최근의 연구 성과를 소개하고자 한다. 먼저 리뷰 분석에 관한 연구와 여행 데이터 분석을 다룬 연구의 방법과 동향에 대해 살펴본 후, 텍스트 마이닝 및 토픽 모델링 방법론의 원리를 요약한다. 그 후 본 연구와 관련이 깊은 딥러닝 방법론, 특히 어텐션 방법론을 살펴본다.

2.1 리뷰 분석

대용량 텍스트 데이터가 도메인별로 증가함에 따라 자동화된 리뷰 분석을 통해 고객의 의견을 분석하는 작업은 거의 필수적인 것으로 인식되고 있다. 도메인마다 특수한 키워드와 패턴이 존재하기 때문에 모두 일괄적으로 같은 방법을 적용할 수는 없지만, 적용되는 방법론의 전체적인 틀은 크게 다르지 않다. 리뷰 분석은 우선 텍스트 데이터 수집 및 전처리 작업, 그리고 전처리한 데이터의 알고리즘 대입, 마지막으로 알고리즘 대입으로 얻은 결과물의 해석이라는 세 가지 단계로 구성된다. 구체적으로 데이터 수집 및 전처리 작업은 불용어를 제거하고 도메인에 최적화된 단어 사전을 사용하여 텍스트 데이터를 정제하는 과정을 거친다. 그 후 연구자는 해당 도메인에 적합한 알고리즘을 사용하여 이전 단계에서 얻은 텍스트 데이터를 분석한 후, 마지막으로 이를 통해 얻은 결과를 해석하는 과정을 거쳐 리뷰 데이터에서 의미 있는 결론을 얻는다.

이 과정은 도메인에 따라 사용될 수 있는 단어 사전이 달라질 수 있다는 점을 제외하면 여러 도메인에서 범용적으로 사용될 수 있다. 이러한 장점은 많은 연구자 및 현업자가 리뷰 데이터 분석을 연구하고 이용하는 데 기여하였고, 이로 인해 최근 전자상거래 기업뿐만 아니라, 호텔 예약, 관광지 리뷰, 등 다양한 도메인에서 더욱 활발한 연구가 이루어지고 있다. 특히 항공승객의 유형별 분석에 관한 연구(남승주, 이형철, 2019), 병원 리뷰 자료를 통한 의료서비스 만족에 관한 연구(이시환 등, 2017), 딥러닝을 활용한 추천 모형 연구(이륜경 등, 2019) 등 다양한 연구가 최근까지 활발하게 이루어지고 있다.

교통 시스템과 인터넷이 발전하면서 관광 산업은 해마다 발전하고 있다. 그 배경에는 저가 항공의 등장, 온라인 예약 시스템의 부상, 그리고 AirBNB와 같은 공유 시스템의 발전이 있었다. 이러한 급속한 관광 산업의 확장세는 기존의 여행사가 제공하는 제품이 아닌, Tripadvisor와 같은 세계 각국의 여행지 정보 및 리뷰를 바탕으로 개인이 직접 정보를 획득하고 여행 코스를 계획하는 이른바 개인 맞춤 여행 시장을 활성화하는 선순환을 이끌었다. Tripadvisor는 세계 각국의 여행지를 대상으로 실제 방문객이 리뷰를 남길 수 있는 형식으로 되어있으며, 유명 여행지의 경우 그 장소에 대한 방문객 리뷰의 수가 약 15만 개 달한다. 이렇듯 여행지 방문객 리뷰의 경우 그 수가 충분히 많을 뿐 아니라 수집이 용이하며, 분석 결과의 활용 가치가 높다는 점에서 많은 연구자들이 연구 주제로 다루어왔다. 이러한 연구의 최근 예로 서울지역의 호텔서비스에 관한 연구(김건, 윤혜정, 2016)와 외국인 관광객의 경복궁에 대한 긍정/부정 리뷰를 분석한 연구(이현주, 2017)를 들 수 있다. 이렇듯 관광객 수가 점차 증가하고 있는 상황에 맞추어 학계에서도 신속한 대응이 이루어지고 있으며, 꾸준히 증가하는 외국인 방문객의 수요에 따라 관광지 활성화를 위한 연구가 더욱 활발하게 수행될 필요가 있다.

2.2 텍스트 마이닝 및 토픽 모델링

인간이 정보를 전달하는 체계에는 그림, 몸짓 등 많은 방법이 존재하지만, 대다수의 의사

소통은 언어, 즉 텍스트에 기반을 두어 이루어진다. 따라서 의사소통의 도구 또는 결과가 방대한 양의 텍스트로 존재하고 있으며, 최근 정보통신 기술 및 하드웨어의 발전으로 트위터, 블로그 등 다양한 서비스를 통해 텍스트 데이터가 매우 빠른 속도로 양산되고 있다. 텍스트 마이닝은 이러한 비정형 텍스트 데이터에 대한 분석을 통해 가치 있는 정보를 추출하기 위한 일련의 방법론을 의미한다.

컴퓨터가 자연어를 처리하는 방식은 인간이 언어를 이해하는 방식과 전혀 다르며, 따라서 텍스트 데이터는 1과 0의 이진값으로 컴퓨터에 저장된다. 보다 구체적으로 컴퓨터는 텍스트를 벡터공간 모델을 이용하여 표현하며(Salton et al., 1975), 이를 기반으로 문서별 단어의 빈도수에 근거한 문서 클러스터링, 문서 자동 분류, 그리고 문서의 주제 식별 및 키워드 추출 등의 후속 분석이 가능하게 된다. 단어를 벡터로 나타내는 과정에서 단순히 출현 빈도만을 반영하는 방법은 많은 한계를 갖고 있으므로, 이를 보완하기 위해 TF-IDF 기반 가중 빈도가 널리 활용되고 있다. TF는 문서 내에서 특정 단어가 등장한 빈도이며, IDF는 이렇게 등장한 단어가 타 문서에서도 등장하는 빈도에 log를 취한 값의 역수이다. TF-IDF 가중 빈도는 TF 값과 IDF 값을 곱하여 산출되며, 이에 따라 특정 문서에서만 높은 빈도로 출현하고 다른 문서에서는 빈번하게 출현하지 않는 단어에 대해 높은 가중치를 부여하게 된다. 해당 방법론은 최근까지도 한국어처리를 위한 방법론으로 활발하게 연구되고 있다(이종화 등, 2019).

일반적으로 토픽 모델링은 TF-IDF 행렬을 이용해 특정 Threshold 이상의 빈도수를 가진 단어를 추출한 후, LDA 알고리즘(Blei et al., 2003)을 통해 각 문서의 토픽을 추출하는 형태로 이루어진다. 구체적으로 본 알고리즘은 먼저 각 문서가 특정한 토픽으로 구성된다고 가정하며 각 단어를 N개의 토픽에 랜덤하게 할당한다. 이후 특정 문서 d에 포함된 단어 중 토픽 t에 해당하는 단어의 비율과 단어 w를 포함하는 모든 문서 중 토픽 t가 할당된 비율을 반복적 실행을 통해 조정함으로써 해당 문서를 구성하는 토픽을 찾아간다. 이를 통해 각 문서는 N개의 토픽 중 확률이 가장 높은 토픽을 배정받으며, 모든 단어는 각 토픽에 대한 가중치에 따라 토픽을 대표하는 단어로 선정된다.

토픽 모델링의 결과를 측정하기 위한 다양한 방법론이 존재하며, 이들 대부분은 토픽 키워드 간 유사성에 기반을 두어 일관성을 평가한다. 일관성 측정을 위한 방법으로는 토픽 키워드를 검출한 후 설문을 통해 일관성을 평가하는 연구(Chang et al., 2009), 키워드 간 PMI (Pairwise Mutual Information) 유사도를 기반으로 토픽의 일관성을 측정하는 연구(Newman et al., 2010), 임베딩된 키워드 유사도를 바탕으로 토픽 일관성을 평가하는 연구(Fang et al., 2016) 등이 있다. 한편 검출된 토픽을 대상으로 군집분석을 수행하여 대표성을 평가하는 방법(Vineet et al., 2014), 중심 안정성(Centric Stability)을 도출하여 최적 토픽 수를 파악 하는 방법(Greene et al., 2014) 등도 수행된 바 있다. 하지만 토픽의 품질은 키워드의 일관성만으로 평가될 수 없기 때문에 다양한 측면에서 토픽의 품질을 측정하기 위한 시도가 이루어질 필요가 있다.

2.3 딥러닝 방법론과 어텐션

로지스틱 회귀, SVM, 의사결정 나무, 신경망 등과 같이 기계학습(Machine Learning)의 한 분야인 딥러닝(Deep Learning)은 개념적으로 여러 층의 신경망을 겹겹이 쌓은 것으로, 여러 층을 거치며 각 뉴런이 복잡한 특성을 파악하고 학습하여 목표 값의 결과를 예측한다. 2000년대에 접어들면서 일반 PC가 32G의 메모리를 사용하는 등 하드웨어는 무어의 법칙을 뛰어넘는 속도로 기하급수적인 성장을 거듭되었고, 그동안 하드웨어적인 제약으로 인해 실현될 수 없었던 딥러닝 방법론이 부상하기 시작했다. 지역적인 특성(Local Features)을 파악하고 이를 통해 뉴런을 활성화하는 CNN(Convolutional Neural Network), 시간 속성을 활용하기 위한 RNN(Recursive Neural Network), 문장의 순서를 고려해 예측값을 생성하는 seq2seq 모델 등 다양한 방법론이 고안되고 있으며, 이들 알고리즘은 기존의 전통적인 머신러닝 알고리즘에 비해 텍스트 분석 분야에서 뛰어난 예측력을 나타내고 있다(Cho et al., 2014; Hochreiter and Schmidhuber, 1997; Kim., 2014).

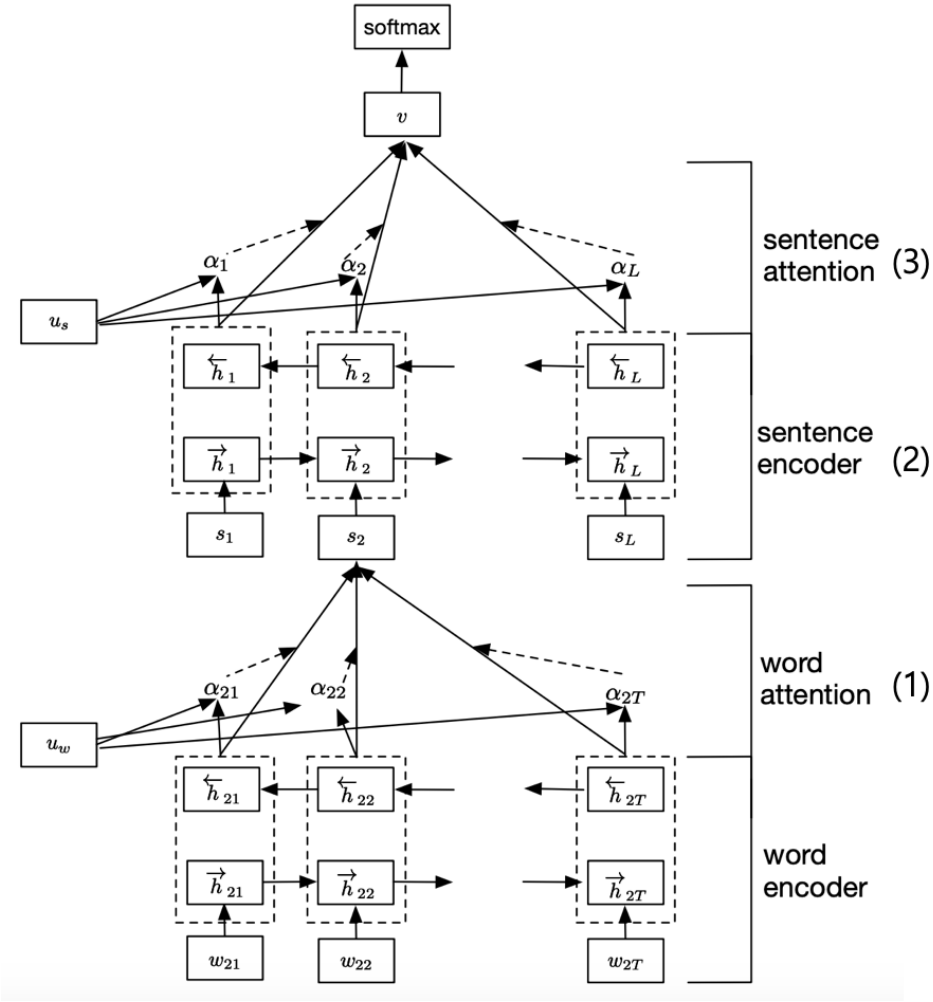
특히 최근에는 seq2seq에 기반을 둔 어텐션 방법론(Bahdanau et al., 2014)이 등장했다. 이는 기존의 seq2seq 디코더가 작전 상태 값만을 고려했던 것과 달리, 인코더의 모든 은닉 상태(Hidden State) 값을 고려하여 다음 결과를 예측함으로써 단어 간 관계를 고려한 학습을 진행하게 된다. 텍스트 분석을 위한 다양한 알고리즘이 어텐션을 응용하여 개선되고 있으며, 본 연구에서는 문서 분류를 통해 단어의 특성을 학습하는 계층적 어텐션 네트워크(HAN: Hierarchical Attention Networks)(Yang et al., 2016)를 활용하여 토픽 모델링의 품질을 개선하고자 한다. HAN은 어텐션 알고리즘을 활용하여 단어별 가중치를 구하고, 이를 바탕으로 문장 벡터를 도출한 후, 이렇게 생성된 문장 벡터가 다시 어텐션 계층과 Softmax 계층으로 전달되어 최종적으로 확률값에 의해 문장을 분류한다. HAN의 전체 구조는 <그림 2>와 같다.

하지만 이러한 개선에도 불구하고 어텐션 메커니즘은 현재 딥러닝 방법론의 근본적인 한계, 즉 여전히 단어의 실제 의미가 아닌 단어의 벡터화를 통한 연산에 기반을 둔다는 한계점을 갖는다. 또한 어텐션 이후 발전된 Transformer 모델(Vaswani et al., 2017), Bert 모델(Devlin et al., 2018)과 비교했을 때 성능이 다소 낮게 나타나는 한계도 보인다. 그럼에도 불구하고 HAN의 경우 과도한 컴퓨팅 자원을 필요로 하지 않으면서도, 비교적 다양한 환경에서 사용할 수 있다는 점에서 본 논문에서는 문장 분류에 HAN을 사용하기로 한다.

본 논문에서는 긍정/부정 리뷰에 대한 분류 학습을 통해 긍정/부정 키워드를 식별하고, 이를 활용하여 토픽 모델링의 결과를 정제하는 방안을 제시한다. 즉 본 연구에서는 긍정/부정 리뷰인 문장에 대한 예측을 수행할 필요가 있으므로, 긍정/부정 키워드 식별을 위해 문장 내 단어 사이의 관계를 고려한 어텐션 기법인 HAN을 사용하여 학습을 진행한다.

Ⅲ. 제안 방법론

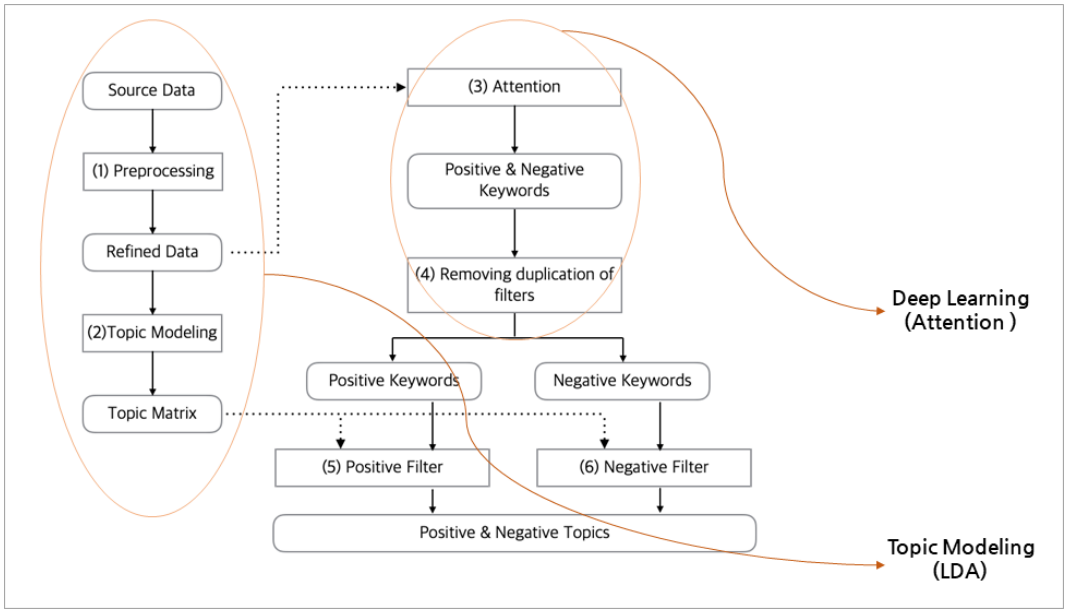
본 장에서는 연구의 전체적인 모형과 각 단



<그림 2> HAN의 전체 구조(Yang et al., 2016)

계의 동작 과정을 소개한다. 이를 위해 먼저 필터링에 사용되는 단어 집단을 “C-word”로 명명하고 이를 이용해 최적의 토픽 키워드를 추출하는 방안을 제시한다. 즉 어텐션 메커니즘을 통해 추출된 C-word를 필터 기준어로 사용하여 LDA를 통해 도출된 토픽 키워드를 정제하게 되며, 전체 연구 모형은 <그림 3>에 제시되어 있다. 그림에서 사각형은 작업을 나타내고, 타원형은 데이터를 나타낸다.

우선 그림의 좌측 부분은 LDA 기반 토픽 모델링을 수행하는 과정으로, 전체 문서를 가장 잘 설명하는 토픽을 찾기 위해 다양한 키워드 수와 토픽 수의 조합에 대해 반복 수행된다. 한편 우측 부분은 본 연구에서 제안하는 방법론의 핵심으로, HAN을 이용하여 각 리뷰의 긍정/부정 분류를 수행하는 과정에서 긍정 키워드와 부정 키워드를 도출한다. 본 연구에서는 긍정/부정 리뷰의 구분에 별점을 사용한다. 즉 높은

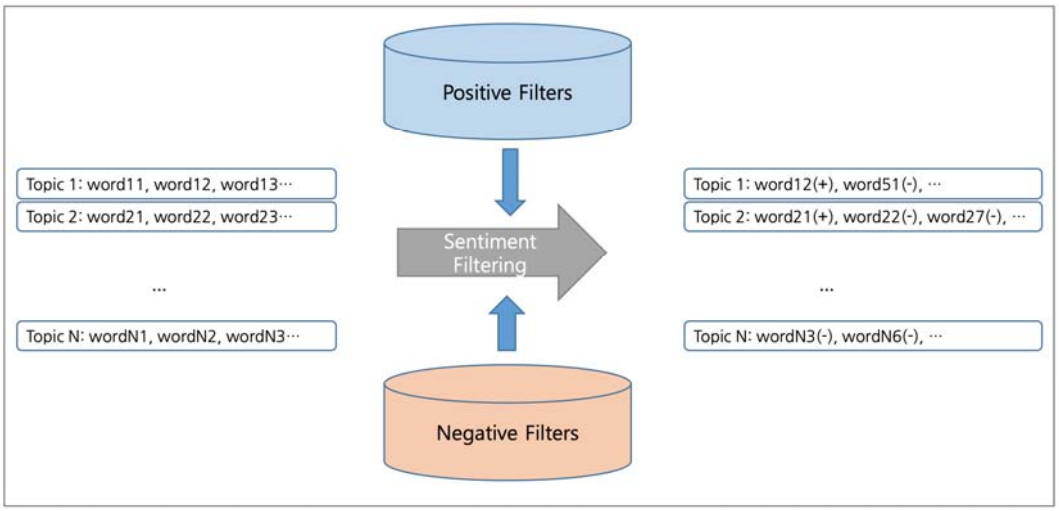


<그림 3> 전체 연구개요

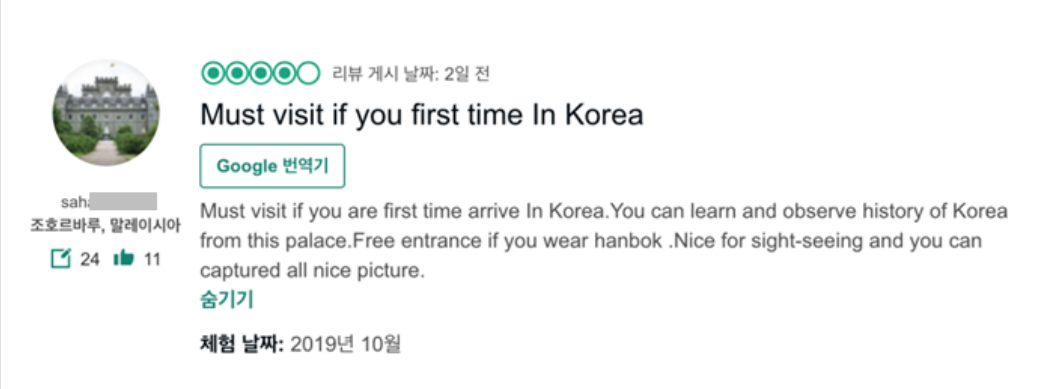
별점과 함께 기록된 리뷰를 긍정 리뷰로, 낮은 별점과 함께 기록된 리뷰를 부정 리뷰로 정의한다. 이렇게 긍정/부정의 타겟을 갖는 문서에 대해 HAN 딥러닝 알고리즘을 사용하여 분류 학습을 진행하며, 학습을 마친 후 모델은 타겟의 레이블(Label)을 예측하는 데 사용된 주요 단어와 가중치 값을 저장하게 된다. 이렇게 도출된 긍정 키워드 집합과 부정 키워드 집합은 방법론의 좌측에서 수행된 토픽 모델링 결과의 필터링에 사용된다.

위의 과정에서 일부 용어는 긍정 키워드 집합과 부정 키워드 집합에 동시에 포함될 수도 있으며, 이러한 현상은 어텐션 알고리즘의 특성에 기인한다. 즉 어텐션 알고리즘에서 각 단어의 의미는 고립된 형태로 규정되는 것이 아니라, 주변 단어들과의 관계에 따라 가변적으로 형성된다. 따라서 동일한 단어라도 주변 단어들에 따라 긍정 또는 부정으로 서로 다르게 사용될 수 있으며, 이러한 단어는 단어 자체만으로 해당 리뷰가 긍정인지 부정인지를 판단하는 단서로 사용되기에는 적합하지 않다. 따라서 긍정 키워드 집합과 부정 키워드 집합에 동시에 포함된 단어들은 상황에 따라 어느 한 쪽 집합 또는 양쪽 집합 모두에서 삭제하고 사용하게 된다. 이러한 일련의 과정을 통해 긍정 필터(Positive Filters)와 부정 필터(Negative Filters)를 생성하게 되고 이들 두 필터를 사용하여 LDA 토픽 모델링에서 도출된 각 토픽별 키워드를 정제함으로써 <그림 4>와 같이 토픽별 긍정 키워드와 부정 키워드를 기술할 수 있다.

본 장에서는 제안 방법론의 전체 개요 및 구조를 소개하였으며, 이어지는 4장에서는 제안 방법론을 실제 데이터에 적용한 분석 과정 및 결과를 소개한다.



<그림 4> 긍정/부정 필터를 사용한 토픽 키워드 정제



<그림 5> Tripadvisor 리뷰 예

Ⅳ. 실험 및 결과

4.1 실험 개요

제안 방법론의 성능을 평가하기 위해, 본 연구에서는 여행 전문 사이트인 Tripadvisor로부터 국내 대표적 관광지 중 하나인 경복궁에 대한 영어권 화자의 리뷰를 수집하여 분석하였다(그림 5). 경복궁의 경우 조선 시대의 궁으로서 상징성도 매우 높을 뿐 아니라, 수도인 서울 시내 중심에 위치하고 있어 접근이 용이하기 때문에 해마다 다수의 관광객이 방문하는 대표적인 명소이다. 또한, 현실적인 측면에서 해당 사이트의 해당 장소는 데이터 수집일 기준 약 6,000개에 달하는 영문 리뷰를 갖고 있으므로 토픽 모델링 및 긍정/부정 분류를 수행하기에 충분한 조건을 갖춘 것으로 판단하였다. <그림 5>와 같은 형태로 표현되는 리뷰로부터 별점,

텍스트 리뷰, 그리고 작성일의 정보를 수집하였으며, 2011년부터 2019년 9월까지의 리뷰 중 영어 텍스트로 작성된 리뷰만을 분석 대상으로 한정하였다.

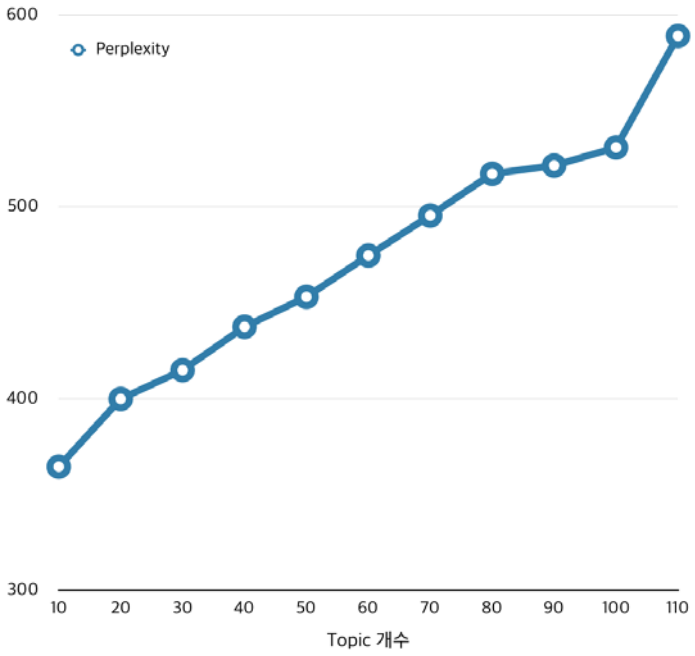
4.2 LDA 토픽 모델링 실험 과정 및 결과

본 절에서는 수집한 데이터에 대해 LDA 기반 토픽 모델링을 수행한 과정과 결과를 소개한다. 토픽 모델링의 전체 과정은 <그림 6>의 의사 알고리즘(Pseudo Algorithm)을 통해 설명되며, 실제 코드는 Python 3.7의 scikit-learn 패키지 기반으로 동작하는 LDA( ) 함수를 이용하여 구현하였다. 전처리 단계에서는 5,920개의 텍스트 데이터를 대상으로 불용어를 제거하였으며, 단어가 어미에 따라 다른 단어로 인식되지 않도록 하기 위해 어간 추출(어미 제거)을 수행하였다. 이러한 전처리 과정에서는 nltk, spacy, 그리고 gensim 패키지를 이용하였으며, 해당 패키지의 불용어 사전 및 lemmatizer를 이용하였다.

다음으로 최적의 토픽 수와 단어 수를 찾기 위해 토픽의 수를 10개부터 100개까지, 그리고 토픽별 단어의 수는 100개 이내로 한정하여 토픽 모델링을 반복 실행하였다. 토픽 모델링에서 최적 토픽 수를 찾는 방안은 이미 많은 연구(Cao and Juan, 2009; Zhao, Moghaddam et al., 2012, Weizhong, et al., 2015)에서 모색되어 왔으며, 대부분의 경우 Perplexity를 참고하여 토픽의 수를 결정할 것을 권장한다. 즉 Perplexity가 낮을수록 토픽 모델링이 적절하게 수행된 것을 암시하는 것으로 해석 가능하다. <그림 7>에서 Perplexity는 토픽의 수가 증가할수록 같이 증가하며, 특히 토픽의 수가 100개 이상일 때 Perplexity의 증가폭이 가파르게 나타남을 확인하였다. 따라서 본 실험의 이후 과정은 10개 이상, 100개 이하의 토픽 수를 갖는 모델에 대해 수행하였다.

```
Procedure TopicModeling{
    lemmatized = [ ]
    rawData = load(TripAdvisor)
    while(rawData){
        dataWithoutSW = removeStopWord(rawData)   // stopword 제거
        lemmatized <- lemmatize(dataWithoutSW)    // stem 추출
    }
    data = tf-idf(lemmatized)                     // 빈도수 기준 상위 1,000개
    dictionary = LDA(data)                        // LDA 토픽 모델링
    saveAsCSV(dictionary)                         // 토픽 모델링 결과 저장
}
```

<그림 6> 토픽 모델링 전 과정의 의사 알고리즘



<그림 7> 토픽 수에 따른 Perplexity의 변화

	Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9
0	excel	citi	palac	palac	place	palac	child	beauti	palac	palac
1	travel	middl	tradit	citi	visit	ticket	high	winter	tour	build
2	restaur	park	hanbok	walk	histori	trip	student	cold	guard	seoul
3	knowledg	easi	wear	station	palac	visit	school	tree	chang	visit
4	transport	late	korean	exit	time	drama	review	spring	guid	korea
5	combin	relax	visit	beij	nice	come	volunt	place	visit	korean
6	famou	modern	free	visit	good	seoul	field	season	time	museum
⋮										
88	away	afford	larg	obviou	castl	queue	hold	short	right	villag
89	great	sunday	tuesday	style	life	look	bring	korea	pretti	countri
90	leav	condit	summer	countri	leav	year	know	cute	littl	backdrop
91	volunt	bright	insid	door	countri	advanc	magnific	bench	know	spend
92	drum	ride	cool	treasur	compound	humid	section	histor	dynasti	major
93	bonu	shop	coupl	carv	long	airport	prior	nearbi	crowd	hous
94	schedul	seren	attir	miss	uniqu	sightse	languag	photogen	stori	templ
95	inner	traffic	color	conveni	atmosphe	sell	work	lunch	mountain	court
96	regular	stair	think	charact	pond	know	spend	kdrama	structur	north
97	umbrella	flag	miss	place	sight	possibl	knowledg	difficult	impress	display
98	replica	uneven	hous	proper	tour	pass	mother	huge	schedul	amaz
99	korea	help	adult	hotel	half	gyeongbo	young	favorit	changdeo	design

<그림 8> 토픽별 주요 키워드 (토픽 수 = 10, 토픽별 키워드 수 = 100)

토픽 모델링의 결과로부터 각 토픽별 주요 키워드를 도출하였다. 전체 실험은 토픽의 수를 10개 ~ 100개로 변화시키며 수행하였으나, 토픽의 수가 지나치게 많은 경우 해석의 어려움이 있으므로 본 절의 이후 부분은 토픽의 수가 10개인 경우에 한정하여 그 결과를 소개한다. 토픽의 수가 10개일 때, 각 토픽에 대해 상위 100개의 단어를 제시한 결과는 <그림 8>과 같다. <그림 8>에 제시된 단어는 이후 어텐션 모델에서 도출된 긍정 필터 및 부정 필터를 사용하여 정제 과정을 거치게 된다.

4.3 Attention을 활용한 C-word 추출

본 절에서는 어텐션 알고리즘을 통해 토픽 키워드의 필터 역할을 하는 C-word, 즉 긍정 필터와 부정 필터를 도출하는 과정 및 결과를 소개한다. 본 과정 역시 Python을 사용하여 구현하였으며, 구체적으로 HAN 방법론을 tensorflow, keras, 그리고 numpy 패키지 등을 이용해 구현하였다. 학습 집합과 검증 집합의 비율은 8:2로 설정하였으며, 5점 스케일의 별점에 대해 1~3점은 부정, 4~5점은 긍정으로 정의하였다. 또한 샘플링을 통해 긍정/부정 리뷰의 수를 각각 1,315로 동일하게 확보하여 실험에 사용하였다.

이렇게 총 2,630개 리뷰에 대한 어텐션 알고리즘 적용을 통해 가중치 기준으로 주요 긍정 키워드와 부정 키워드를 도출하였다. 다만 긍정 키워드와 부정 키워드 수의 균형을 유지하기

긍정 어휘	긍정 가중치	부정 어휘	부정 가중치
photogenic	0.985752314	disappoint	0.99668793
gorgeous	0.971521499	budget	0.957566414
clean	0.966659996	mediocre	0.914040331
colourful	0.961567307	prefer	0.891366567
photography	0.941451341	commercial	0.87838486
mountain	0.933033767	today	0.822150662
recommend	0.932245393	waht	0.818071006
attractive	0.929316844	bosintang	0.79304768
scenic	0.928852241	understatement	0.786879202
highlight	0.925433984	uninteresting	0.773299495
surreal	0.893317364	pride	0.756872957
strongly	0.891865132	minimalist	0.748928654
memorial	0.875974995	superstition	0.739987575
romantic	0.857583487	creepy	0.617106
⋮			

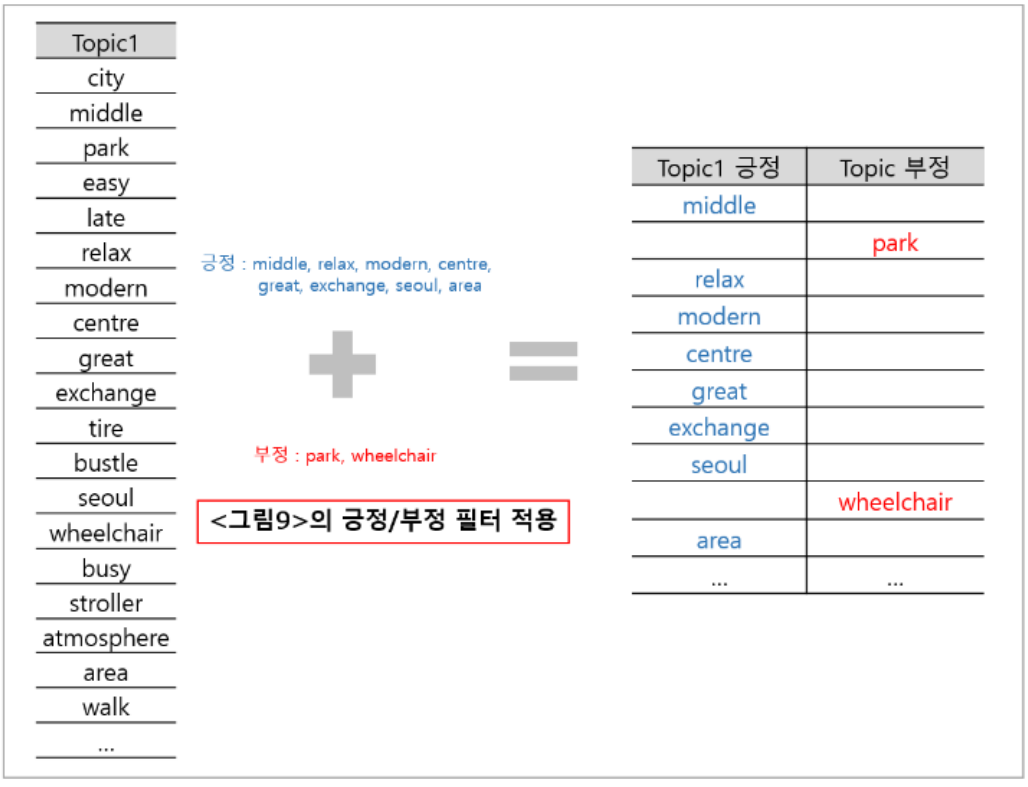
<그림 9> 긍정 필터 및 부정 필터

위해, 긍정 키워드의 경우 가중치 0.15 이상의 단어 385개를 추출하였고 부정 키워드의 경우 보다 완화된 기준인 가중치 0.04 이상을 적용하여 240개의 단어를 추출하였다. 이렇게 도출된 C-word, 즉 긍정 필터와 부정 필터의 일부가 <그림 9>에 나타나있다. 이후 과정은 <그림 9>의 용어 중 긍정/부정의 양 필터에 속한 일부 단어를 제외하고 긍정 키워드 382개, 부정 키워드 168개를 사용하여 수행한다.

4.4 토픽 키워드 필터링

본 절에서는 실험의 마지막 단계, 즉 <그림 8>의 토픽 키워드에 대해 <그림 9>의 긍정/부정 필터를 적용하여 각 토픽 키워드를 긍정/부정 관점에서 정제하고 식별하는 과정 및 결과를 소개한다. 전체 과정은 <그림 10>과 같으며, <그림 10>의 좌측은 정제 이전의 Topic1 키워드를, <그림 10>의 우측은 정제 이후의 Topic1 키워드를 나타낸다. 정제 과정에서 “city”, “easy”, “late” 등의 단어는 제거되었으며, “middle”, “relax”, “modern” 등의 단어는 Topic1을 구성하는 긍정 키워드로, “park”, “wheelchair”는 Topic1을 구성하는 부정 키워드로 식별되었음을 알 수 있다.

<그림 10>과 같은 정제 과정을 전체 토픽의 키워드에 대해 긍정/부정 키워드를 식별한 결과가 <표 1>에 나타나있다. <표 1>은 각 토픽의



<그림 10> Topic 1 키워드의 정제 과정 및 결과

주제와 함께 긍정/부정 키워드를 요약하여 제시하고 있다. <표 1>은 토픽 모델링을 통해 도출한 <그림 8>의 토픽에 대해 어텐션을 통해 구축한 <그림 9>의 긍정/부정 필터를 적용하여, 각 토픽의 주제와 함께 긍정/부정 키워드를 요약하여 제시하고 있다.

<표 1> 전체 토픽 키워드에 대한 긍정/부정식별 결과

토픽	긍정/부정 키워드
Topic 1. 경복궁 주변의 음식점	**[긍정]** travel, combine, superb, sight, short, serene, speak, fully, miss, organize, meet, tour, absolute, enjoy, love, history, profession, require, learn, experience, forget, visit, vast, join, discovery, seoul, present, great, bonus, regular, korea **[부정]** time, folklore, basic, replica
Topic 2. 경복궁의 주변 환경	**[긍정]** middle, relax, modern, centre, exchange, area, seoul, area, love, sunny, peace, stroll, surround, attract, worth, place, piece, incredible, monday, allow, escap, maintain, visitor, beauty, ground, wait, locate, large, nice, land, serene, help **[부정]** park, wheelchair, pond, wall, push, process, uneven
Topic 3. 전통 복장 체험	**[긍정]** tradition, visit, dress, place, photo, change, beauty, ground, experience, seoul, great, ceremony, nice, huge, area, korea, enjoy, spend, love, recommend, watch, crowd, worth, come, feel, main, cafe, need, visitor, definite, locate, outfit, amaze, allow, attract, forget, able, large, cool, couple, color, miss **[부정]** time, picture
Topic 4. 경복궁과 중국의 자금성 비교	**[긍정]** visit, nice, number, seoul, locate, large, area, stop, early, main, temple, remind, ground, short, allow, worth, change, touristy, choose, style, country, miss, place **[부정]** Beijing, train, instead, straight, version, door
Topic 5. 한국의 역사와 문화	**[긍정]** place, visit, history, nice, beauty, seoul, great, culture, enjoy, love, amaze, learn, huge, area, recommend, thing, definite, worth, korea, photo, feel, need, preserve, come, experience, crowd, spend, know, ancient, clean, plenty, miss, peace, stroll, maintain, able, change, relax, nature, temple, surround, watch, vast, stay, tranquil, castle, country, sight, tour **[부정]** picture, spot, weather, pond

Topic 6. 한국 드라마와 관광객	**[긍정]** trip, visit, drama, come, seoul, need, early, crowd, tour, hear, worth, group, feel, place, recommend, special, purchase, beauty, attract, thing, shot, help, scene, weekend, turn, minute, photo, foreign, huge, korea, week, able, reserve, history, november, visitor, limit, watch, year, airport, sightsee, know **[부정]** time, picture, sign, board, single, pass
Topic 7. 학생들의 자원봉사	**[긍정]** important, history, place, profession, photograph, teenage, offer, festival, receive, regret, total, holiday, present, story, trip, weekend, organize, group, bonus, vast, absolute, monday, ancient, like, corner, strongly, know, section, language, spend **[부정]** uneven, simply, money, replica, artifact
Topic 8. 계절별 경복궁의 특색	**[긍정]** beauty, spring, place, season, wish, visit, love, nice, mountain, amaze, gorgeous, surround, absolute, week, rainy, maintain, cafe, incredible, stun, festival, awesome, photo, outdoor, apart, charm, stop, allow, memory, serene, short, korea, photogen, kdrama, huge, favorite **[부정]** winter, weather, background, time, toilet, picture, simply, hand
Topic 9. 수문장 교대식에 대한 반응	**[긍정]** tour, change, visit, ceremony, ground, history, seoul, beauty, worth, spend, korea, place, watch, recommend, inform, miss, enjoy, great, huge, definite, nice, check, large, love, culture, main, catch, language, thing, join, explain, experience, offer, speak, area, plan, learn, color, colour, photo, help, amaze, group, minute, know, crowd, story, mountain **[부정]** -
Topic 10. 우리나라와 일본의 건물 비교	**[긍정]** seoul, visit, korea, ground, large, beauty, area, main, year, restore, history, tradition, locate, reconstruct, feel, worth, attract, modern, visitor, color, place, huge, mountain, ancient, surround, style, import, actual, great, country, spend, temple, north, display, amaze **[부정]** rebuild, time, expect, quarter, pond, japan, wall

각 토픽별 긍정/부정 키워드를 살펴보면 전체 토픽 모두 긍정 키워드가 많이 도출되었으며 부정 키워드는 상대적으로 적게 나타남을 알 수 있다. 심지어 Topic 9. 수문장 교대식에 대한 반응 토픽의 경우 부정 키워드가 전혀 포함되지 않았는데, 실제로 경복궁을 찾는 외국인 관광객들의 반응이 가장 긍정적인 체험이 수문장 교대식인 것으로 알려져 있다.

<표 1>과 같이 긍정/부정 키워드가 식별된 토픽 모델링의 경우, 주요 주제에 대해서 긍정적인 측면과 부정적인 측면을 구분하여 해석할 수 있다는 장점을 갖는다. 예를 들어 Topic 5.

한국의 역사와 문화의 경우 culture, history, ancient, nature, tranquil 등과 같은 긍정 키워드가 도출되었다. 이를 통해 관광객들이 경복궁에서 한국의 역사와 문화를 느낄 수 있을 뿐 아니라, 조용하면서도 자연과 조화를 이루는 경복궁에 대해 긍정적으로 평가함을 알 수 있었다. 한편 picture, spot 등 부정 키워드도 식별되었는데, 실제 리뷰 원문 분석 결과 이는 사진을 찍는 장소가 다소 한정적이라는 반응에 따른 결과인 것으로 해석된다.

본 장에서 수행한 실험의 결과, LDA 토픽 모델링을 통해 도출된 토픽 키워드에 대해 어텐션 알고리즘을 통해 획득한 긍정/부정 키워드 필터를 적용함으로써, 주요 토픽을 구성하는 요소를 긍정/부정으로 구분하여 파악할 수 있음을 확인하였다.

Ⅴ. 결 론

본 연구에서는 긍정/부정 리뷰에 대한 분류 학습을 통해 긍정/부정 키워드를 식별하고, 이를 활용하여 토픽 모델링의 결과를 정제하는 방안을 제시하였다. 구체적으로 본 연구에서는 LDA 기반 토픽 모델링을 수행하여 토픽 키워드를 추출하고, 이와 별도로 딥러닝 기법 중 하나인 어텐션 기반 학습을 수행하여 긍정/부정 키워드를 추출한 후, 긍정/부정 키워드를 필터로 사용하여 토픽 키워드를 정제하는 방안을 제시하였다. 제안 방법론의 실무 적용 가능성을 확인하기 위해 대표적인 여행 전문 사이트인 Tripadvisor로부터 국내 대표적인 관광 명소인 경복궁에 대한 영문 리뷰 약 6,000건을 수집하여 분석을 수행하였으며, 그 결과 제안 방법론에 의해 주요 토픽을 구성하는 긍정 키워드와 부정 키워드를 적절하게 잘 식별함을 파악하였다.

본 연구의 학술적 기여는 다음과 같다. 기존 토픽 모델링은 토픽을 구성하는 주요 키워드들을 제시함으로써 토픽의 주요 내용을 파악하는 데 도움을 주지만, 토픽을 구성하는 각 키워드가 긍정적인 의미로 포함되었는지, 반대로 부정적인 의미로 포함되었는지를 확인할 수는 없다는 한계를 갖고 있다. 이와 달리 본 연구에서는 토픽 모델링과 딥러닝을 결합하여 토픽 키워드를 긍정과 부정 요소로 구분하여 제시하였다는 점에서 기존 연구와의 차별성이 인정될 수 있다. 또한 본 연구에서는 텍스트 분석 분야에서 가장 활발하게 연구가 이루어진 LDA 토픽 모델링과, 최근 딥러닝 분야에서 최신 알고리즘으로 관심이 높아지고 있는 어텐션 메커니즘을 접목시킨 연구라는 점에서 그 기여가 인정된다. 토픽 모델링의 경우 기본적으로 주요 토픽, 즉 높은 빈도로 출현하는 단어를 발견하는 것을 목표로 수행되며, 어텐션의 경우 타켓 클래스의 식별, 즉 본 연구의 경우 긍정과 부정 리뷰를 식별하는 데 중요하게 사용되는 단어를 발견하는 것을 목표로 수행된다. 따라서 두 기법을 결합한 본 연구의 경우 주제를 구성하는 주요 단어 중 긍정/부정의 식별력을 갖고 있는 단어만을 선별하여 결과로 제시할 수 있으며, 향후 이러한 방식의 다양한 시도가 학계에서 활발하게 이루어질 수 있을 것으로 기대한다.

이러한 학술적 기여에도 불구하고, 본 연구의 기여는 실무적 측면에서 더욱 강조될 수 있을 것으로 기대한다. 본 연구는 관광지에 대한

리뷰 분석을 통해, 관광객들의 주요 반응에서 긍정적 요소와 부정적 요소를 구분하여 해석할 수 있다는 가능성을 보였다. 이를 통해 각 관광지의 장점을 파악하여 강화하고, 동시에 단점을 파악하여 보완하는 전략을 수립하기 위해 본 연구에서 제안한 방법론이 효과적으로 활용될 수 있을 것으로 기대한다.

하지만 이러한 학술적/실무적 기여에도 불구하고 본 연구는 몇 가지 한계를 갖는다. 우선 많은 양의 학습 데이터를 어텐션 알고리즘의 특징으로 인해 최소한 수천 건 이상의 리뷰를 갖는 관광지에 대해서만 분석이 가능하다. 따라서 향후 전이 학습(Transfer Learning) 등의 적용을 통해 소량의 데이터에 대해서도 제안 방법론을 적용할 수 있는 방식으로 연구의 확장이 이루어질 필요가 있다. 또한 본 논문의 2장에서 전술한 바와 같이 토픽 키워드의 품질을 일관성 측면에서 평가하기 위한 시도는 다수 이루어진 바 있지만, 토픽의 활용성 측면에서의 품질 평가는 아직 이루어지지 못했다. 본 연구는 토픽 키워드를 긍정과 부정 요소로 식별하여 토픽 키워드의 활용성을 증대시킬 수 있는 방안을 제시하였지만, 이를 통한 실제 효과에 대한 정량적인 검증이 이루어지지 않았다. 따라서 향후 연구에서는 설문조사 등 다양한 방법을 통해 제안 방법론의 성과에 대한 더욱 엄밀한 평가가 이루어져야 한다. 또한 본 논문에서 실험에 사용한 데이터의 경우 긍정적인 평가가 부정적인 평가에 비해 상대적으로 많았기 때문에, 분석을 통해 도출한 긍정 키워드의 수가 부정 키워드의 수에 비해 많게 나타났다는 특징을 갖는다. 향후 연구에서는 분석 대상 사례의 긍정 평가와 부정 평가의 비율을 감안하여 실험을 수행할 필요가 있다.

참고문헌

김건, 윤혜정, “토픽모델링을 활용한 서울지역 호텔서비스에 대한 고객인식의 변화 분석,” 서비스경영학회지, 제17권 제3호, 2016, pp. 217-231.

남승주, 이현철, “LDA 토픽 모델링을 활용한 항공승객 유형 별 특성 분석,” 경영과학, 제36권 제3호, 2019, pp. 67-85.

이륜경, 정남호, 홍태호, “딥러닝을 이용한 온라인 리뷰 기반 다속성별 추천 모형 개발,” 정보시스템연구 제28권 제1호, 2019, pp. 97-114.

이시환, 조아람, 이훈영, “온라인 병원 리뷰자료의 Latent Dirichlet Allocation 분석을 활용한 의료서비스 만족 요인에 관한 연구,” 서비스경영학회지, 제18권 제5호, 2017, pp. 23-44.

이종화, 이문봉, 김종원. “TF-IDF 를 활용한 한글 자연어 처리 연구,” 정보시스템연구 제28권 제3호, 2019, pp. 105-121.

이현주, “빅데이터를 활용한 경복궁 방문 경험 분석,” 대한관광경영학회지, 제32권 제2호, 2017, pp. 297-318.

Andrzejewski, D., and Zhu. X., “Latent dirichlet allocation with topic-in-set knowledge,” Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing, Association for Computational

Linguistics, 2009, pp. 43-48.

Bahdanau, D., Cho, K., and Bengio, Y.. "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.

Blei, D. M., Ng, A. Y., and Jordan, M. I., "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, Vol. 3, 2003, pp. 993-1022.

Blei, D. M., and Lafferty, J.D., "Dynamic Topic Models," In Proceedings of the 23rd International Conference on Machine Learning, June 2006, pp. 113-120.

Cao, J., Xia, T., Li, J., Zhang, Y., and Tang, S., "A density-based method for adaptive LDA model selection," *Neurocomputing*, Vol. 72, No. 7, 2009, pp. 1775-1781.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M., "Reading tea leaves: How humans interpret topic models," In Advances in neural information processing systems, 2009, pp. 288-296.

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y., "On the properties of neural machine translation: Encoder-decoder approaches," arXiv preprint arXiv:1409.1259, 2014.

Cho, K., Courville, A., and Bengio, Y., "Describing multimedia content using attention-based encoder-decoder networks," *IEEE Transactions on Multimedia*, Vol. 17, No. 11, 2015, pp. 1875-1886.

Cui, G., Lui, H. K., and Guo, X., "The effect of online consumer reviews on new product sales," *International Journal of Electronic Commerce*, Vol. 17, No. 1, 2012, pp. 39-58.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K., "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

Fang, A., Macdonald, C., Ounis, I., and Habel, P, "Using word embedding to evaluate the coherence of topics from twitter data," In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, July 2016, pp. 1057-1060.

Greene, D., D.O. Callaghan, and P. Cunnungham, "How Many Topics? Stability Analysis for Topic Models," ECMLPKDD'14 Proceedings of the 2014th European Conference on Machine Learning and Knowledge Discovery in Databases-Volume Part I, 2014, pp. 498-513.

Hochreiter S., and Schmidhuber J., "Long short-term memory," *Neural computation*, Vol. 15, No.9, November 1997, pp. 1735-80.

Landauer, T. K., Foltz, P. W., and Laham, D.,

"An introduction to latent semantic analysis," *Discourse processes*, Vol. 25, No. 2, 1998, pp. 259-284.

Liu, B., Dai, Y., Li, X., Lee, W. S., and Yu, P. S., "Building Text Classifiers Using Positive and Unlabeled Examples," Proceedings of the 3rd IEEE International Conference on Data Mining, 2003, pp. 179-188.

Kim, Y., "Convolutional Neural Networks for Sentence Classification," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), arXiv preprint arXiv:1408.5882, 2014.

Moghaddam S., and Ester M., "On the design of LDA models for aspect-based opinion mining," InProceedings of the 21st ACM international conference on Information and knowledge management, October 2012, pp. 803-812.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T., "Automatic evaluation of topic coherence," In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, June 2010, pp. 100-108.

Rubin, T. N., Chambers, A., Smyth, P., and Steyvers, M., "Statistical topic models for multi-label document classification," *Machine learning*, Vol. 88, No. 1, 2012, pp. 157-208.

Salton, G., Wong, A., and Yang, C. S., "A vector space model for automatic indexing," *Communications of the ACM*, Vol. 18, No. 11, 1975, pp. 613–620.

Tasci, S., and Gungor, T., "LDA-based keyword selection in text categorization," In 2009 24th International Symposium on Computer and Information Sciences, September 2009, pp. 230-235.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... and Polosukhin, I. "Attention is all you need," In Advances in neural information processing systems, 2017, pp. 5998-6008.

Vineet, M., R.S. Caceres, and K.M. Carter, "Evaluating Topic Quality Using Model Clustering," 2014 IEEE Symposium on Computational Intelligence and Data Mining, 2014, pp. 178-185.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E., "Hierarchical attention networks for document classification," In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, June 2016, pp. 1480-1489.

Zaremba, W., Sutskever, I. and Vinyals, O., "Recurrent neural network regularization," arXiv preprint arXiv: 1409.2329, 2014.

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., and Zou, W., "A heuristic approach to determine an appropriate number of topics in topic modeling," *BMC bioinformatics*, Vol. 16, No. 13, December 2015, Available : https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-16-S13-S8/.

최 병 설 (Choi, Byeong-Seol)

평생교육원에서 수학사 학위를 취득하였으며, 현재 국민대학교 비즈니스IT전문대학원 석사과정에 재학 중이다. 주요관심 분야는 Text Mining, Deep Learning, Data Modeling 등이다.

김 남 규 (Kim, Namgyu)

현재 국민대학교 비즈니스IT전문대학원장 및 경영정보학부 교수로 재직 중이다. 서울대학교 컴퓨터 공학과에서 학사 학위를 취득하고, KAIST 테크노경영대학원에서 Database와 MIS를 전공하여 경영공학 석사 및 박사 학위를 취득하였다. 한국 지능정보시스템 학회 부회장, 한국정보기술응용학회 부회장, 한국경영학회 상임이사, 한국경영정보학회 이사, 한국인터넷 정보학회 이사를 역임하였다. 주요 관심 분야는 Text Mining 및 Data Mining, Deep Learning, Data Modeling 등이다.

<Abstract>

Topic Modeling with Deep Learning-based Sentiment Filters

Choi, Byeong-Seol · Kim, Namgyu

Purpose

The purpose of this study is to propose a methodology to derive positive keywords and negative keywords through deep learning to classify reviews into positive reviews and negative ones, and then refine the results of topic modeling using these keywords.

Design/methodology/approach

In this study, we extracted topic keywords by performing LDA-based topic modeling. At the same time, we performed attention-based deep learning to identify positive and negative keywords. Finally, we refined the topic keywords using these keywords as filters.

Findings

We collected and analyzed about 6,000 English reviews of Gyeongbokgung, a representative tourist attraction in Korea, from Tripadvisor, a representative travel site. Experimental results show that the proposed methodology properly identifies positive and negative keywords describing major topics.

Keyword: Big Data, Deep Learning, Hierarchical Attention Networks, Review Analysis, Text Analytics, Topic Modeling,

* 이 논문은 2019년 12월 2일 접수, 2019년 12월 14일 1차 심사, 2019년 12월 18일 게재 확정되었습니다.